Creating an Academic Hotbot: Final Report of the University of Michigan OAI Harvesting Project

John Wilkin,
Primary Investigator
jpwilkin@umich.edu

Kat Hagedorn,
Project Librarian
khage@umich.edu

Mike Burek,
Project Programmer
mburek@umich.edu

Contents

Summary

The University of Michigan OAI harvesting project, Creating an Academic Hotbot, concluded its work in December, 2002. The first phase involved identification of potential metadata sources and early exploration. The second phase of work involved installing and deploying the harvester software developed by the University of Illinois, Urbana Champaign (UIUC), and testing and collaborating with UIUC on fixes to the software. With the stabilized software, OAI-enabled repositories from various sources were harvested. A formal end-user Web interface for searching the harvested metadata was created and deployed at this point. The third phase of work involved creating and running a Web survey and intensive in-house end-user interviews to test the successes and limitations of the current search interface. The fourth phase of work involved a second round of end-user interviews, search log analysis, research into relevancy of results, and the development of a revised search interface.

OAIster can be found online at http://www.oaister.org/, with more than a million records available from over 120 sources. OAIster staff members have incorporated their software work on the interface into the DLXS distribution process, so that part of OAIster is now available through Open Source licensing and has been installed at some of the nearly thirty DLXS sites around the world. Further dissemination of UM’s software work (especially XSLT-based transformation routines) will be made available during the coming year.

Background

In early 2001, the University of Michigan Library proposed to the Mellon Foundation the establishment of a broad, generic, information retrieval resource. This service, to be built through a collaboration that relies on the University of Illinois’s proposed metadata harvesting tool, would provide users with access to information about publicly available digital library objects where those objects were declared through the OAI protocol. Michigan proposed focusing on metadata for information resources that are publicly accessible and have no access restrictions, and have a corresponding Web-based digital representation (e.g., this would not include the metadata records for slides when the slides cannot be accessed through the Web). As proposed, the service also intended to encompass as broad a collection of resources as possible (i.e., with no subject parameters), and was to be accessible to the entire Internet community, without bounds. The middleware that Michigan proposed using for the access system was to be made available freely to other institutions for implementation as they see fit, e.g., to develop subject-based OAI-enabled repositories.

Phase One: Identification of Metadata and Early Exploration (December 10, 2001 - February 28, 2002)

The University of Michigan project was awarded funding from the Mellon Foundation in June, 2001. Phase one of the project was largely exploratory and ran in parallel to the development of the project’s metadata harvester at the University of Illinois, Urbana Champaign (UIUC). Michigan began a process of posting positions and hiring for its two project positions in late summer. Both staff members, Kat Hagedorn (project librarian) and Mike Burek (project programmer), began work in December. Hagedorn began contacting potential sources of metadata, and both Hagedorn and Burek began familiarizing themselves with the OAI landscape—those working on the OAI protocol, those developing OAI-enabled repositories, those becoming service providers who would harvest those repositories, those building tools to provide and harvest, and those interested in the open archives, digital libraries and free scholarship movements in general. Both staff members became proficient quickly, and during this time were developing tools and creating early implementations that would shape their later efforts. For example, Burek created a simple harvester that gathered many of the records that were processed in the prototyping efforts. Hagedorn developed proficiency with the DLXS Bibliographic Class (BibClass) software, mounted a number of collections, and used her growing knowledge of OAI and Dublin Core to help shape the DLXS "broker" (metadata exposure) software.

Phase Two: Delivery of a Service Based on UIUC Tools (March 1, 2002 - June 30, 2002)

Use of the OAI protocol for this project involved the following steps:

1. Harvesting Dublin Core (DC) encoded metadata in XML format;
2. Using locally-created tools to transform that DC metadata into BibClass encoded metadata; and
3. Indexing and serving this metadata to users through an interface that uses the XPAT search engine, which Michigan licenses through the Digital Library eXtension Service (DLXS).

In mid-2002, Michigan released its service, called OAIster, and began distributing the middleware for OAIster through DLXS. A report on those activities follows.

Harvesting and Transformation

Rather than develop its own harvester, Michigan relied on a parallel development effort at UIUC. UIUC developed this harvester initially using Microsoft technology, but adapted it as a Linux-based Java harvester (relying on MySQL) to be used by Michigan in its Unix-based environment. Michigan is currently running version 2.0B4 of the Java harvester as created by UIUC.

It was also necessary to transform the records from DC to our own native format. The transformation tool Michigan developed is written in Java and relies on XSLT to transform DC records to BibClass records. Steps in this process (carried out by the Java-based software) include:

1. Collecting individual records that are presented by the harvester in directory trees into large files ready for XSLT transformation;
2. Removing records that do not have digital objects associated with them;
3. Normalizing the contents of the DC element Resource Type[1];
4. Adding institution information to each record;
5. Converting UTF-8 to ISO8859-1 for the purposes of indexing and retrieval;
6. Transforming (via XSLT) DC records into BibClass;
7. Counting records and providing quality of data feedback.

[1] Using a normalization table, DC Resource Type values such as "book" and "paper" are transformed to the normalized value "text," and values such as "illustration" and "picture" are transformed to the normalized value "image." The table was manually created from a retrieval of unique DC Resource Type values among all harvested records.

The transformation involved mapping from DC to BibClass elements, illustrated in Table 1.

Table 1. Dublin Core to BibClass element mapping.

Original Elements	**Example Value**	**BibClass Element**	**Displayed as…**
OAI Element			
identifier	oai:VTETD:etd-92398-135228	ID attribute for A (i.e., complete record) element	for internal use
datestamp	1998-10-23	DT attribute for A (i.e., complete record) element	for internal use
DC Element			
title	Estimating Exposure and Uncertainty for Volatile Contaminants in Drinking Water	K	Title
creator	Sankaran, Karpagam	L	Author/ Creator
contributor	Mary Leigh Wolfe	M	Contributor
subject	Civil and Environmental Engineering	SU	Subject
description	The EPA recently completed a major study to evaluate exposure and risk associated with a primary contaminant, radon and its progeny in drinking water (EPA, 1995). This work …	AA	Note
publisher	Virginia Polytechnic Institute and State University	T	Publisher
date	1998-10-23	YR	Year
type	text	TYPE	Resource Type
format	application/pdf	FMT	Resource Format
identifier	http://scholar.lib.vt.edu/theses/available/etd-92398-135228/	URL	URL
language	en	LANG	Language
rights	I hereby grant to Virginia Tech or its agents the right to archive and to make available my thesis or dissertation …	X	Rights

All metadata values are displayed “as is,” without modification. At present, the DC Source element is mapped to a BibClass Note element, and the DC Relation element is not mapped or displayed. The Author display field was re-visited in September, 2002, based on input from users, and was changed to Author/Creator, for better clarification.

One of the results of the work performed thus far has been a refinement of methods for aggregating and disseminating metadata from extremely heterogeneous repositories. Repositories vary in several ways, including formats (e.g., text, video), academic levels (e.g., graduate student theses, peer-reviewed articles), and topics (e.g., physics, religious studies). And, of course, the repositories vary significantly in the quality of their metadata, including their use of Dublin Core. The transformation methods that Michigan developed have taken this variation into account, and can handle these processes in extremely robust ways. The Michigan transformation tools are in the process of being made public for use by others.

Michigan also continues to encounter and work through several problems associated with the OAI metadata harvesting protocol. Some examples follow:

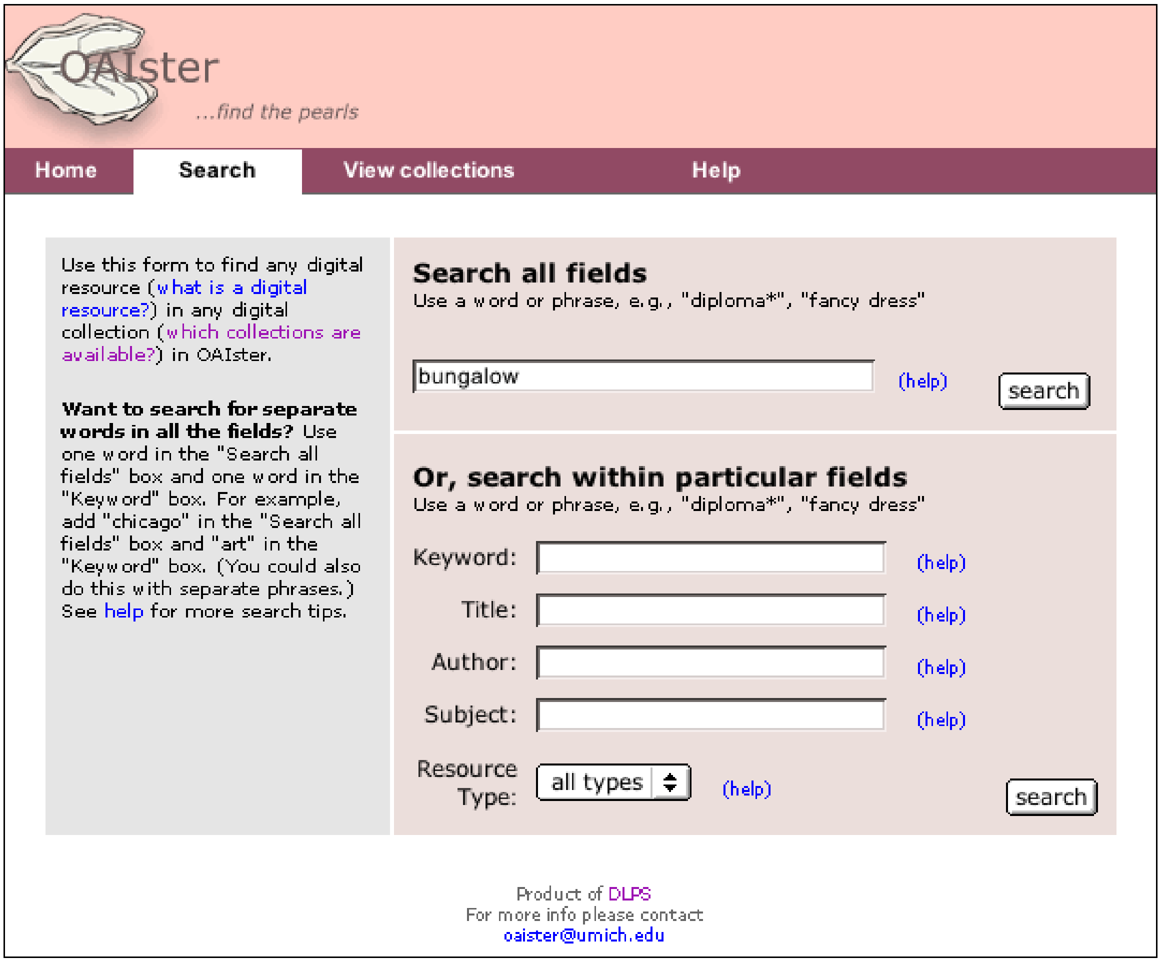
- Records are duplicated at provider sites for a variety of reasons, particularly because of the role of metadata aggregators who both harvest and expose metadata.
- A provider may be exposing records whose digital objects are restricted to licensed users; currently, there is no standardized method of indicating restrictions on access to digital objects (e.g., an OAI protocol element with binary "restricted" and "unrestricted" values).
- Much of the metadata available for harvesting is not valid XML (e.g., does not use appropriate UTF-8 encoding), and so produces unusable records.
- Scheduling harvesting can be challenging, as long harvesting efforts can often end up overlapping, and thus causes problems with memory-intensive, concurrent processes.

We expect that a majority of these issues will be more tractable as OAI harvesting and exposure becomes more widespread and tools are refined.

Search and Retrieval

Michigan launched the first formal OAIster search interface on June 28, 2002 with 274,062 records from 56 repositories. Figure 1 is a screenshot of the original search interface.

Figure 1. Original OAIster search interface.



Use this form to find any digital resource (what is a digital resource?) in any digital collection (which collections are available?) in OAIster.

Want to search for separate words in all the fields? Use one word in the "Search all fields" box and one word in the "Keyword" box. For example, add "chicago" in the "Search all fields" box and "art" in the "Keyword" box. (You could also do this with separate phrases.) See help for more search tips.

Search all fields

Use a word or phrase, e.g., "diploma*", "fancy dress"

bungalow

Or, search within particular fields

Use a word or phrase, e.g., "diploma*", "fancy dress"

Keyword:

Title:

Author:

Subject:

Resource Type: all types

Product of DLPS
For more info please contact
oaister@umich.edu

The results of user testing ran concurrently with OAIster development (see Phase Three, below), and informed the current search interface. The OAIster interface represents a significant departure from the default BibClass interface, while keeping the underlying architecture. Changes to BibClass came in four primary areas:

- The search interface (http://www.oaister.org/cgi/b/bib/bib-idx?c=oaister;page=simple) was altered slightly from the DLXS simple full-text search interface (e.g., http://www.hti.umich.edu/cgi/t/text/text-idx?page=simple&c=umhistmath). The structure of the search page display remained the same (apart from coloring and font changes to reflect the OAIster image), but did not include elements designed specifically for full-text searching (i.e., searching in a particular region). The OAIster interface also included limiters in addition to author and title (e.g., resource type, subject) in order to research whether these would be useful for searching bibliographic collections.
- Staff altered the display of results to limit the number of records displayed to ten. Often, harvested records were lengthy in particular because of the descriptive notes (see Phases 3 and 4, below, for assessment of the note field). We also modified the display of the header information on the results page to make it potentially easier to read, and created more white space on the page to maximize comfort for scanning.
- Individual record formatting was changed to emphasize the field/value display. We wanted to associate the field label (metadata element) more closely with the value to increase readability, hence the right-hand justification of the label and the choice of color to make the display more clear.
- Staff modified the method for author/title sorting, such that titles with no authors were not interspersed among the results, but showed up at the end of the results. This new method is being incorporated into several of the DLXS classes.

We received anecdotal evidence on the success and limitations of our initial release. Several users emailed to indicate that they were disturbed by our decision to limit the number of records that are retrieved. Others found the interface difficult to use because of the small font or because it was difficult to determine how to formulate multi-word searches. See Phase 4, below, for our solutions to these problems. Nevertheless, users were pleased that there was a service like this available. We have received positive comments, such as:

- “Splendid service, and I will promote it widely!”
- “An excellent resource—I have already made good use of it twice this morning!”
- “I’m not up [at the Media Union, a “remote” campus building] often. But with these kinds of resources why would I need to go?!”
- “I think it’s a great service—and a wonderful site to use to illustrate the power of the OAI effort.”

Many comments came from researchers in the digital library environment who had been informed in the initial announcements. The service had not yet been widely promoted within academia at that point. A press release was created in December, 2002 and picked up by the University News and Information Service, the School of Information Newsletter, and the University Record (a weekly publication) to date. Contacts with the Chronicle of Higher Education have been made, and we hope for inclusion in that publication, as well as other academic trades, soon.

Phase Three: First Assessment and Experimentation
(March 1, 2002 - June 30, 2002)

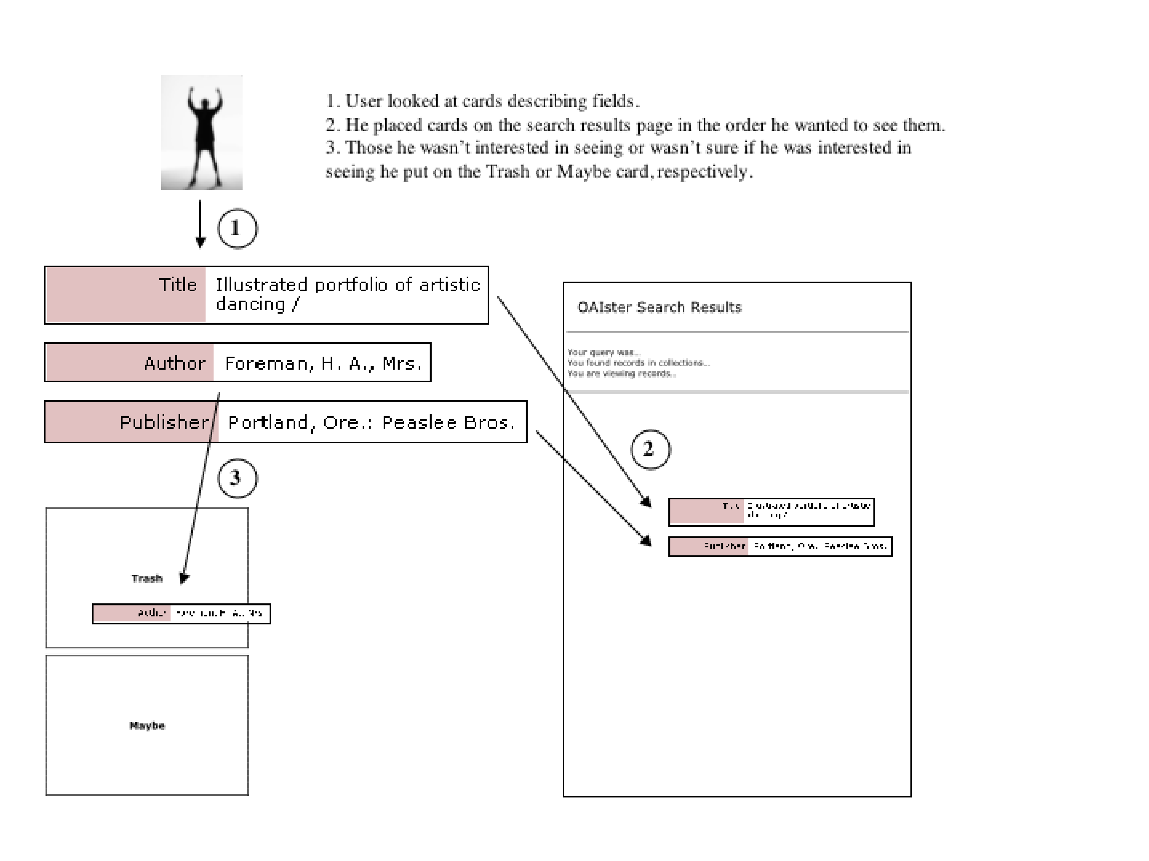
Project staff devoted considerable time to assessment and evaluation of possible interfaces. They designed an online survey that could assist them in determining what users might want from a system like OAIster, and particularly what sorts of digital resources users might be interested in when working online, what they look for but are not able to find, and some of the problems they run into when looking for information online. A summary, along with the original questions, is available at http://www.oaister.org/o/oaister/surveyreport.html. We received 591 responses over the month that the survey was open. Some of the most interesting findings were:

- A majority of respondents indicated they were most interested in online journals and reference materials when they went online to look for information.
- Additionally, respondents indicated that these were the digital resources they were unable to find online.
- Top problems that respondents noted when looking for information online included not retrieving the resource itself, not finding older materials, having rights problems, not knowing of a comprehensive service for finding resources, and the ever-present problems with searching.
- Several potential features of a service like OAIster were of interest to respondents, including looking for resources in a certain subject area, using a service that is continually updated, and searching the full text of the resource.

Based on the initial survey, staff worked to design a potential user interface. The interface incorporated functionality drawn from the several years of work behind the interfaces of the DLPS collections, and showcased the aggregated metadata we harvested. OAIster designs started on paper and were subsequently discussed, changed and reviewed again. Usability staff tested the provisional design with 9 users, one-on-one, in front of a computer. These users ranged from experts in searching digital collections to novices. They were asked to review mockups of the provisional interfaces in a web browser, including a search page, a page where they would select a specific repository to search in, and a search results page.

Questions were open-ended, and testing also included early prototyping. For example, users were asked “What do you think this page is for? What do you think you can do here?” and “How would you go about finding images of Monet’s “Water Lilies” series using this page?” Usability staff also asked users to tell them if there was anything they found lacking on the mockups by writing their answers on an actual print of the mockup, which seemed to garner useful results. At the end of the session, users performed a “paper prototyping” exercise—indicating the metadata elements or fields they would want to see on a search results page by placing cards containing those elements on a blank “interface”, as illustrated in Figure 2 below.

Figure 2. Paper prototyping exercise.



1. User looked at cards describing fields.
2. He placed cards on the search results page in the order he wanted to see them.
3. Those he wasn't interested in seeing or wasn't sure if he was interested in seeing he put on the Trash or Maybe card, respectively.

1
Title
Illustrated portfolio of artistic dancing /
Author
Foreman, H. A., Mrs.
Publisher
Portland, Ore.: Peaslee Bros.
3
Trash
Maybe
OAIster Search Results
Your query was...
You found records in collections...
You are viewing records...
2

The results of the in-person user testing helped reveal what was needed for the original interface. Some of the most interesting results were:

- Appropriate labeling was needed in the interface. Some users had difficulty defining digital library terms such as “collections,” “repositories,” and “digital resources.” Labeling was addressed more fully for the revised interfaces (see Phase 4, below).
- Some processes were difficult for users to understand. The repository selection page was quite cumbersome for most users, so this added functionality was removed from the initial interface release. (See Phase 4, below, for discussion on why it was not added for the revision of the interface.)
- Suggestions for better arrangements of useful information on the page. When users wrote what they felt was lacking on a printout of the actual mockup, they tended to agree on placement of elements on pages. For example, users were fairly clear about what they wanted in the results summary area and where that should be placed on the page.
- The paper prototyping exercise was designed to find out what fields users wanted to see on long and short versions of a results page. Results indicated that users didn't want a shortened results page—they were interested in viewing as many fields as possible at one time.

Phase Four: Further Assessment, System Revision, and Dissemination (July 1, 2002 - November 30, 2002)

The OAI protocol has increased in popularity in the year and a half since it was developed. In one month's time (June to July, 2002), over ten new OAI-compliant repositories were registered on the official Open Archives Initiative site.[2] By the end of November, 2002 OAIster was serving nearly a million records (we have since topped a million) from over 120 repositories. Our statistics have stabilized to the point where over 2000 searches are made per month. (We expect this to increase after the press release is disseminated more widely.)

In the final phase of the project, we conducted valuable research on relevancy of results that could inform a digital library community in need of such data. In addition, we performed more user testing and analyzed our search log statistics to assist us in building a revised search and search results interface.

Also included in the project is dissemination of the software. The OAIster software has already been distributed as part of DLXS and is now available to approximately 30 DLXS institutions around the world. The University of Amsterdam, one of two European DLXS sites, will begin making available a portal with OAI-accessible metadata, and has preliminary plans to build on the OAIster interface and functionality. The Universities of Sydney and South Africa are also DLXS sites, and we hope to work with them to mount similar services. Many more institutions can benefit from the XSLT-based transformation tools we have developed, and which we will soon distribute.

Second Round of User Assessment

A second round of user testing was performed in September and October of 2002. We tested all users one-on-one, as in the first round of testing, but split the testing among in-house users and remote users. Users were asked to look at the live version of OAIster, instead of mockups as before, but as before were asked to review certain pages (namely, the search page and the search results page) and answer a variety of questions.

We presented the users with several scenarios and asked questions based on these scenarios. For example, we asked users to imagine themselves as a reference librarian. A patron approaches and asks for information on women in farming in the early 1900s. We then asked questions such as "What is your first instinct for search terms on the search page?" and "On the results page, can you tell how the results are ordered?"

The most interesting results were:

- Most users know some part of what they are looking for, but also take the time to browse contextually and/or look for information that is associated with what they have found.
- Common post-find actions that were desired were printing, bookmarking, downloading, and incorporating the digital object into another document or file.
- In order to determine whether a result is useful or not, users look most often at the search terms they used (these are highlighted in bold and red in the results), and at both the Title and Notes fields. Because users seemed to focus on the notes field and preferred more

[2] Many repositories are moving to the latest version of the protocol (2.0). To be listed as registered on the official OAI site, they must be compliant to the 2.0 protocol. The list of registered repositories is the first place we look for new repositories.

field information to less (from the first round of testing), it was decided not to offer a shortened version of the results page that would let users link to “more information” about the record they were looking at.

- Surprisingly, searching by institution (i.e., for a particular repository) was not desired for users. Even sorting by institution was not a priority.
- Users understood that OAIster was designed for academia, that they could use OAIster to find trustworthy, authentic digital objects.
- While they had some trouble with the original search interface (particularly in how to use the different search boxes), they liked the search results page very much, especially in its use of white space, and placement of field labels and values.
- Searching seemed to be not so much subject specific (which we were having them do) as format specific (which is what they wanted to do). In other words, instead of limiting by further subject terms, they would rather limit by the format of the digital object, e.g., text, image, etc.
- Users like to sort, by date, by format, although not by institution. Multiple sorting options seem not to be a hindrance, but a benefit.

Revised Interface

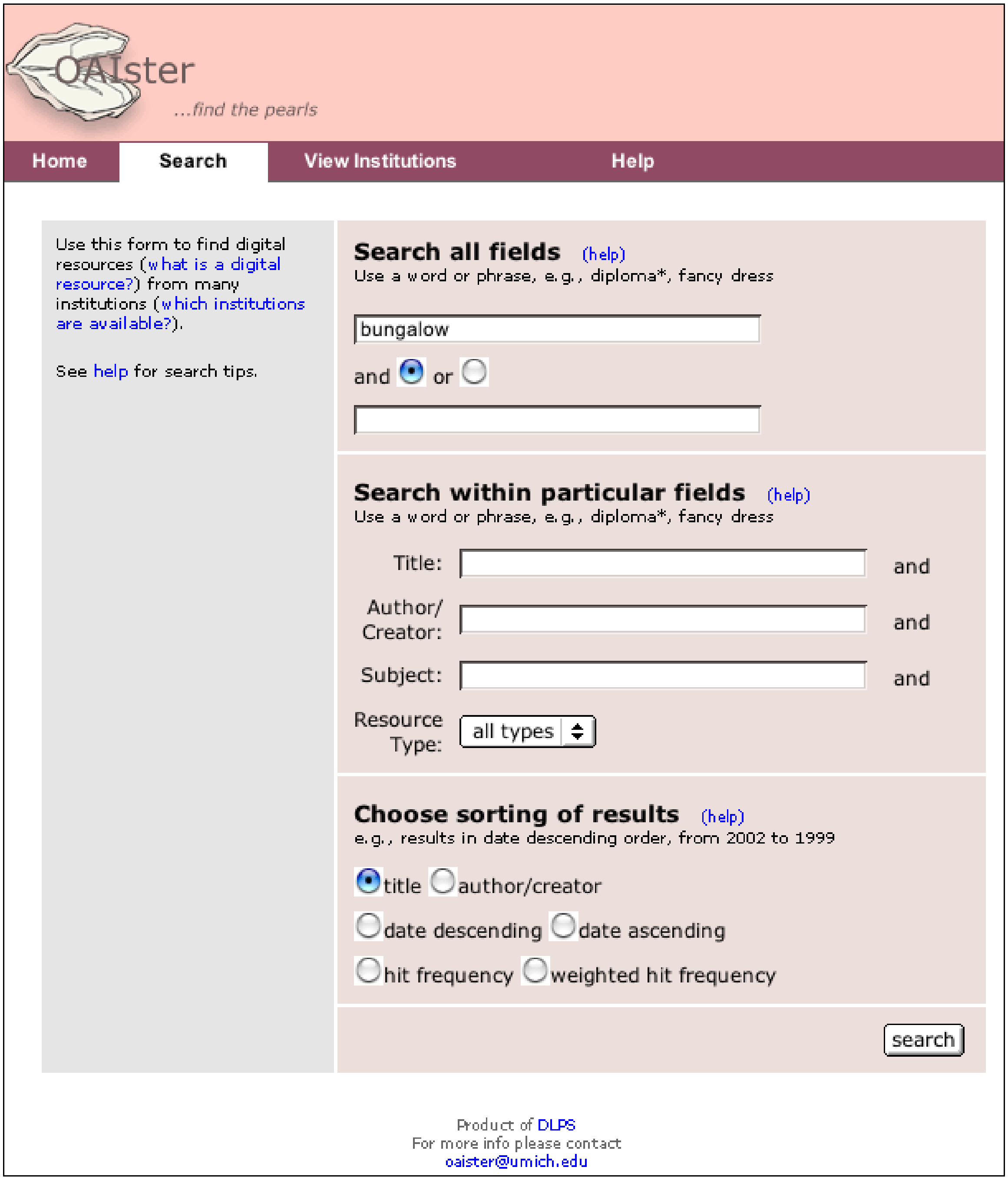
OAIster staff members made several enhancements between the end of June and the end of September of 2002. After the second round of user testing, the results were used to make significant modifications to the search interface and the search results interface:

- Included limited Boolean search capability. Both AND and OR were added, with NEAR and NOT excluded, to keep the interface as simple as possible. In the original search interface, users could use Boolean logic, but only after considered study of the interface, or a thorough perusal of the help file.[3]
- Explanatory language was re-thought, particularly for the sorting options, so that users could more easily understand how to use them in their search.
- Provided the opportunity for users to select their sorting option directly from the search interface, instead of only from the results interface. Two relevancy-related sorting methods were also created and made available at this point (see the relevancy of results research, below).
- Allowed users to revise the search they just made.
- Made it possible for users to view all the records they found. Due to time limitations during the initial release, we were not able to make the interface changes necessary to accommodate this. The interface is currently in an interim state in which users can view all the retrieved records but are not able to sort them. A modification of the code and interface is being worked on, as time permits, to provide a stable solution to this.

These revisions are illustrated in Figures 3 and 4 below.

[3] The Boolean AND operator could be used in the simple search interface by adding a term in the first search box (e.g., “aquaculture”) and then a term in the “Keyword” box (e.g., “fish”). These boxes both search all indexed metadata elements. In this roundabout fashion, which admittedly was poorly designed in the initial interface, users could search more than one word, not as a phrase.

Figure 3. Revised OAIster search interface.



OAIster
...find the pearls

Home | Search | View Institutions | Help

Use this form to find digital resources (what is a digital resource?) from many institutions (which institutions are available?).

See help for search tips.

Search all fields (help)
Use a word or phrase, e.g., diplomа*, fancy dress

bungalow

and ● or ○

Search within particular fields (help)
Use a word or phrase, e.g., diploma*, fancy dress

Title: and

Author/ Creator: and

Subject: and

Resource Type: all types

Choose sorting of results (help)
e.g., results in date descending order, from 2002 to 1999

● title ○ author/creator

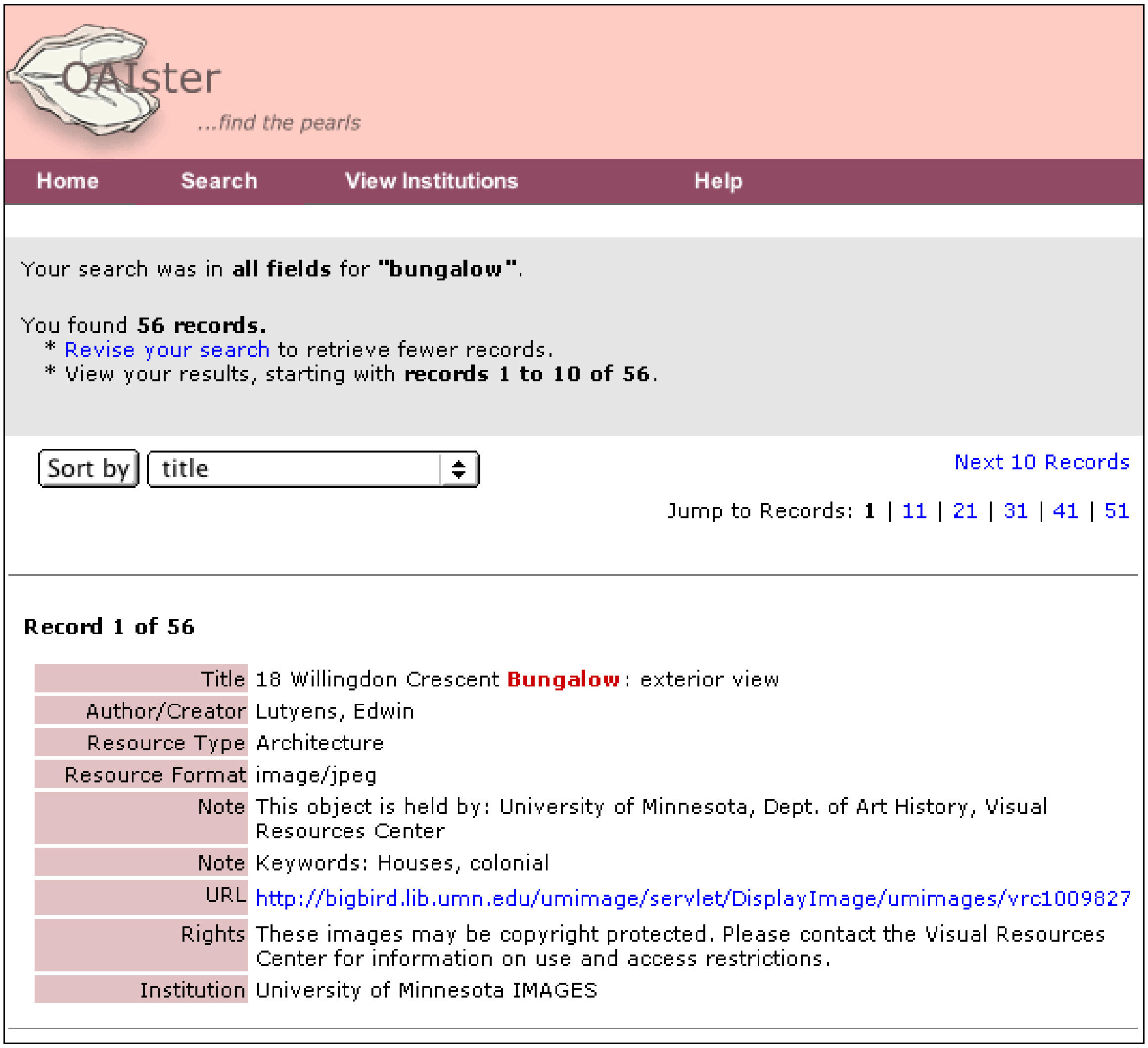
○ date descending ○ date ascending

○ hit frequency ○ weighted hit frequency

search

Product of DLPS
For more info please contact
oaister@umich.edu

Figure 4. Revised OAIster search results interface.



Your search was in **all fields** for **"bungalow"**.

You found **56 records.**
* Revise your search to retrieve fewer records.
* View your results, starting with **records 1 to 10 of 56**.

Next 10 Records

Jump to Records: **1** | 11 | 21 | 31 | 41 | 51

Record 1 of 56

Title	18 Willingdon Crescent **Bungalow**: exterior view
Author/Creator	Lutyens, Edwin
Resource Type	Architecture
Resource Format	image/jpeg
Note	This object is held by: University of Minnesota, Dept. of Art History, Visual Resources Center
Note	Keywords: Houses, colonial
URL	http://bigbird.lib.umn.edu/umimage/servlet/DisplayImage/umimages/vrc1009827
Rights	These images may be copyright protected. Please contact the Visual Resources Center for information on use and access restrictions.
Institution	University of Minnesota IMAGES

Statistical Trends

Analysis of OAIster day-to-day search logs shows some trends in use of the search service. The following table (Table 2) illustrates (by using two example months during which the search service was up and running) how many searches were made and which institutions were using the search service most heavily. (The University of Michigan is included as a comparison value.)

Table 2. Analysis of day-to-day search logs: total searches and top searching institutions.

Types of Statistics	**Example Months**	
	July	*September*
Total Number of Searches	8321	2536
Top Five Institutions Using OAIster	Boston College = 94 State University of New York, Buffalo = 55 Glasgow University = 32 Northern Arizona University = 27 Rijksuniversiteit Gent, Gent, Belgium = 26 (University of Michigan = 317)	University of Wisconsin, Madison = 101 New York University = 62 University of North Carolina at Chapel Hill = 46 The University of Melbourne, Melbourne, Australia = 41 University of Southampton, England (tied) = 31 University of Rochester (tied) = 31 (University of Michigan = 258)

While the number of searches has decreased since launch, the usage has increased by a small percentage across institutions that use OAIster most often. The institutions using OAIster vary widely (across all months), which could be a factor of insufficient, inconsistent marketing of the service. Overall, we expect OAIster to become more useful to the general public as more repositories are added and the type of material included becomes more diverse.

Before changes were implemented to the search service in the fall of 2002, we looked at the usage of Boolean searching (in its rudimentary form) and search limiters[4]. After the search interface changes, we looked at this again. We selected three days from each set and calculated the percentage of searches in which Boolean AND was used, and the percentage of time that search limiters were used. This is illustrated in Table 3 below.

Table 3. Analysis of day-to-day search logs: percentage use of Boolean logic and search limiters.

Sample Dates Selected (in 2002)	**Percentage Use of Boolean AND**	**Percentage Use of Search Limiters**
07/01, 07/18, 07/30	2.2% (20 out of 905 total searches)	8.1% (73 out of 905 total searches)
11/18, 12/03, 12/19	12.0% (19 out of 158 total searches)	26.6% (42 out of 158 total searches)

Interestingly, although the number of searches between the first and second set of sample dates decreased, the percentage use of the Boolean AND operator and search limiters increased. This could be evaluated in a number of ways, including that users had an easier time understanding the revised search interface, and that we were attracting users more familiar with advanced search techniques by the later dates.

Search terms that users entered varied widely, and uncovered a number of interesting issues:

- Misspellings, e.g., “blue swede shoes.” In many search engines, users can expect to have their misspellings accounted for by the system.
- Multiple words strung together, e.g., “east detroit halfway.” This seems to indicate that users expect to search the system as they would any web search engine, with multiple

[4] Users can limit their search using a number of fields: Title, Author, Subject and Resource Type. Resource Type is the normalized field. An example of the use of a limiter is entering “duisburg” in the first box on the page and “grimm” in the Author box.

words searched separately instead of as a phrase. We hope that the more obvious method for Boolean AND searching alleviates this problem.

- Limiters used after trying one word or phrase, e.g., first search = "bibliographic instruction"; second search = "bibliographic instruction" with "sutherland." in the Author limiter box. This may indicate that users are retrieving too many records for them to peruse the first time. If they retrieve over 1000 records, these will be shown to them unsorted, which may be too hard to manage.

Relevancy of Results Research

Included in the project proposal were research questions focused on determining whether relevancy of results is important for users in an academic search service. We undertook to answer these questions by first designing two new methods of sorting called "hit frequency" and "weighted hit frequency," which were included in the revised interface. Both of these methods can arguably be defined as providing users with more relevant digital objects at the top of their results list.

"Hit frequency" sorting counts the number of instances of the words and phrases entered and orders them from highest count to lowest count. "Weighted hit frequency" also counts instances, but gives more weight to instances of words and phrases in certain fields, for instance, a hit in the title field is counted as "100," subject as "40" and notes as "20." The records display a score (e.g., 210) if weighted hit frequency is chosen.

We wanted to test whether these new methods of sorting were useful. Because we had already performed two substantial rounds of user testing, we decide to combine this question with our other research question—how would specialist librarians determine the "best answers" for a particular subject search? We reasoned that "best" answers were the most "relevant" answers, so we developed a small testing environment to compare how librarians, as users of the service, chose relevant objects versus how the new sorting methods chose relevant objects.

Specialist librarians were each asked to perform searches in the subject matter they were experts in. We chose individuals with science expertise, namely chemistry, engineering, math, and physics. Two individuals were chosen for each subject matter, so we could compare the answers to each question. For instance, we asked both chemistry librarians to assume that a faculty member in molecular biochemistry wanted to find 5 relevant documents on this subject matter. They were asked to assume that they had thought about this and decided to perform a search in the revised search interface using the terms "polymer" and "acid." We then had each librarian choose the 5 most relevant digital objects from the results (of no more than 25). At the same time we performed this search ourselves, using the "hit frequency" and "weighted hit frequency" sorting methods.

While the results were interesting, we should state one caveat first. The specialist librarians found it difficult to determine what was relevant, in particular because they felt the subject matter was too broadly stated for the terms they were asked to enter, and that they were not allowed to conduct a reference interview to more specifically define what the faculty member wanted. These are both valid issues and pointed out that this was a study that may better be performed in a more real-life setting.

With this caveat in mind, our most interesting results were:

- There was little reason to assume that our sorting methods were better or worse than a person at finding relevant documents. To illustrate this point, in the following table (Table 4), for the electrical engineering search ("electrical" in the first box and "circuitry" in the second box) we compare librarians and sorting methods in terms of the relevant digital objects they chose.

  Table 4. Comparison of relevant digital objects chosen by librarians and sorting methods for the electrical engineering search.

Comparable searches between librarians and sorting methods for relevant digital objects	**Correspondence of digital objects chosen**
Librarian 1 compared with "weighted hit frequency"	0.40 (2/5)
Librarian 2 compared with "weighted hit frequency"	0.60 (3/5)
Librarian 1 compared with Librarian 2	0.40 (2/5)

  This example indicates that the "weighted hit frequency" sorting method chose some of the same relevant digital objects as the librarians did. One could say that this means the sorting methods worked well, but the percentage of correspondence between librarians and the "weighted hit frequency" sorting method is low. (Correspondence was similar for the "hit frequency" sorting method.)
- The highest correspondence came from the librarians alone. For the physics subject search, the librarians had a 1-to-1 correspondence with each other in terms of relevant digital objects chosen. In this case, the sorting methods did do more poorly, as each method had only a 0.40 correspondence with the librarians.
- There was no indication that "weighted hit frequency" performed better than "hit frequency." The digital objects chosen as relevant by each of the sorting methods for a particular search varied, so we can assume that if users wish to utilize either sorting method, it is better to have them both so users have more ways to view their results.

Naturally, more testing needs to be done on whether users find these sorting methods useful or not. Specialist librarians are potential end-users of the system, but they did not use the sorting methods themselves in the above test. However, it seems that these methods are not without some validity, based on our findings. After further tests, it would be important to work on tweaking the sorting methods or building more appropriate ones that might better mirror the methods that users employ to determine relevancy.

Postscript

As discussed in our proposal, the University of Michigan service has been integrated into the base-funded production operations at Michigan, and will be sustained as long as OAIster provides a relevant and meaningful service. We have contacted Larry Page at Google about cooperation between OAIster and Google, perhaps integrating OAIster results into Google's specialized searches. We plan to continue development of OAIster as well, and when possible will devote resources to the following:

- Provide high-level topical (or similar) browsing capabilities, perhaps drawing on the "sets" functionality in the OAI protocol.
- Work to eliminate or otherwise process duplicate records.
- Normalize more elements, such as DC Language.
- Collaborate with other projects that could benefit from using OAIster, e.g., giving researchers the ability to find digital objects while developing their courses online in a course tools environment.
- Target particular audiences within the research community (e.g., we have begun to develop partnerships with communities such as South Asian bibliographers to provide specialized portals as a subset of OAIster).

As OAI becomes more popular, there will be more opportunities for different ways to aggregate metadata—by topic, by format, by audience, by geographic entity. We hope that there will be more numerous and more varied service providers in the near future, thus enriching the type of information found through OAIster.

Appendix: Budget

The following table (Table 5) contains the budget information for this project. Overall, the project ran close to the proposed budget. Deviations from the budget are described below.

- Salaries were a little higher than budgeted due to the opportunity to hire more experienced staff.
- The University of Michigan Libraries covered the overage detailed in the budget table as cost share.
- Supplies and travel were less than anticipated because they were supplemented or covered by University of Michigan sources.

Table 5. OAIster budget table.

OAISTER BALANCE SHEET															
Expenses														**Projected**	
	Dec 2001	Jan 2002	Feb 2002	Mar 2002	Apr 2002	May 2002	Jun 2002	July 2002	Aug 2002	Sept 2002	Oct 2002	Nov 2002	Dec 2002	Jan 2003	Total
Salaries	$ 2,730.16	$ 13,126.99	$ 9,000.00	$ 9,395.28	$ 9,395.28	$ 10,185.85	$ 10,750.04	$ 10,493.51	$ 11,072.56	$ 9,766.70	$ 9,744.02	$ 9,794.11	$ 3,035.85	-	$ 118,490.35
Benefits	$ 271.65	$ 3,259.31	$ 2,435.96	$ 2,529.60	$ 2,529.62	$ 2,720.08	$ 2,681.82	$ 2,668.56	$ 2,706.99	$ 2,592.38	$ 2,586.03	$ 2,332.46	$ 740.67	-	$ 30,055.13
Computer Services & Supplies	-	$ 1,560.00	$ 126.45	$ 518.90	-	-	-	-	-	-	-	-	-	$ 3,894.82	$ 6,100.17
Travel	-	-	-	-	$ 520.00	$ 719.22	$ 12.00	-	-	-	$ 88.00	-	-	-	$ 1,339.22
Indirect Cost	-	-	-	-	-	-	$ 3,360.97	$ 3,290.52	$ (6,651.49)	-	-	-	-	-	-
General Expenses	-	-	$ 9.00	-	$ 19.95	-	-	-	-	-	$ 105.00	$ 75.87	-	-	$ 209.82
Total	$ 3,001.81	$ 17,946.30	$ 11,571.41	$ 12,443.78	$ 12,464.85	$ 13,625.15	$ 16,804.83	$ 16,452.59	$ 7,128.06	$ 12,359.08	$ 12,523.05	$ 12,202.44	$ 3,776.52	$ 3,894.82	$ 156,194.69
Total Revenues Received	$ 150,000.00														
Interest Received	$ 4,189.62														
Expenses to Date (Dec. 2001 – Dec. 2002)	$ (152,299.87)														
Projected Expenses (Jan. 2003)	$ (3,894.82)														
Projected Balance, End of Project	$ (2,005.07)														
BUDGET STATUS															
	Budgeted	**Projected Actual**													
Salaries	$ 104,732.00	$ 118,490.35													
Benefits	$ 29,324.96	$ 30,055.13													
Computer Services & Supplies	$ 9,000.00	$ 6,100.17													
Travel	$ 5,000.00	$ 1,339.22													
Indirect Cost	-	-													
General Expenses	-	$ 209.82													
	$ 148,056.96	$ 156,194.69													